A Bayesian adaptive marker-stratified design for molecularly targeted agents with customized hierarchical modeling

Yong Zang[1,2*], Beibei Guo[3], Yan Han[1], Sha Cao[1,2] and Chi Zhang[2,4]

1. Department of Biostatistics, Indiana University

2. Center for Computational Biology and Bioinformatics, Indiana University

3. Department of Experimental Statistics, Louisiana State University

4. Department of Medical and Molecular Genetics, Indiana University.

*For correspondence: Yong Zang, Department of Biostatistics, Indiana University, 410 W 10th Street, Indianapolis, IN, 46075, US. Email: zangy@iu.edu

Abstract

It is well known that the treatment effect of a molecularly targeted agent (MTA) may vary dramatically, depending on each patients' biomarker profile. Therefore, for a clinical trial evaluating MTA, it is more reasonable to evaluate its treatment effect within different marker subgroups rather than evaluating the average treatment effect for the overall population. The marker-stratified design (MSD) provides a useful tool to evaluate the subgroup treatment effects of MTAs. Under the Bayesian framework, the beta-binomial model is conventionally used under the MSD to estimate the response rate and test the hypothesis. However, this conventional model ignores the fact that the biomarker used in the MSD is in general predictive only for the MTA. The response rates for the standard treatment can be approximately consistent across different subgroups stratified by the biomarker. In this paper, we proposed a Bayesian hierarchical model incorporating this biomarker information into the consideration. The proposed model uses a hierarchical prior to borrow strength across

This is the author's manuscript of the article published in final edited form as:

Zang, Y., Guo, B., Han, Y., Cao, S., & Zhang, C. (2019). A Bayesian adaptive marker-stratified design for molecularly targeted agents with customized hierarchical modeling. Statistics in Medicine, 38(15), 2883–2896. https://doi.org/10.1002/sim.8159

different subgroups of patients receiving the standard treatment and therefore improve the efficiency of the design. The prior informativeness is determined by solving a "customized" equation reflecting the physician's professional opinion. We developed a Bayesian adaptive design based on the proposed hierarchical model to guide the treatment allocation and test the subgroup treatment effect as well as the predictive marker effect. Simulation studies and a real trial application demonstrate that the proposed design yields desirable operating characteristics and outperforms the existing designs.

KEY WORDS: Adaptive design, Bayesian method, Biomarker, Clinical trial, Hierarchical modeling, Marker-stratified design, Molecularly targeted agents, Subgroup treatment effect.

1 Introduction

Molecularly targeted agents (MTAs) have revolutionized the way that physicians treat patients by enabling them to select treatments tailored to patients' specific biomarker profile[1]. MTAs are expected to be more effective and less toxic than conventional chemotherapy and radiotherapy because they block the growth of cancer cells by identifying and attacking specific functional biomarker while sparing normal cells at the same time[2]. Upon their development, MTAs have been used to treat breast cancer, multiple myeloma, prostate cancer and other types of cancer[3].

The biomarker plays a key role in the MTAs development and personalized treatment selection. According to different functionalities in the process of treatment selection and disease diagnosis, biomarkers can be broadly categorized as predictive or prognostic[4,5]. A predictive biomarker is one that is used to foretell the differential efficacy of a particular therapy based on the presence or absence of the biomarker, e.g., only patients whose tissues highly express the biomarker are expected to respond favorably to a specific MTA. A prognostic biomarker is one that separates a population with respect to the risk of a specific outcome, such as disease progression, in the absence of treatment or despite receiving a non-targeted standard treatment. A typical example of the prognostic marker is the clinical tumor stage as patients with severe stage of cancer tend to respond worse than patients with early stage of cancer regardless of what treatment they receive. Predictive and prognostic biomarkers have different clinical utilities and the one an MTA targets for is in general a predictive marker. For example, the MTA tamoxifen targets for the biomarker estrogen receptor (ER), which is a predictive marker in the sense that tamoxifen is effective only for patients with overly-expressed ER[6].

Due to the biological mechanism of MTAs, the performance of such agents varies remarkably depending on patients' biomarker profile. Therefore, the treatment effects of MTAs

are generally evaluated within different subgroups stratified by patient's predictive marker status[7]. The marker-stratified design (MSD) is arguably one of the most popular biomarker-guided clinical trial designs to evaluate the subgroup treatment effect of MTAs and has been used to conduct a number of clinical trials[8,9,10]. The MSD stratifies patients into a marker-positive subgroup and a marker-negative subgroup according to the patients' predictive marker status, and then randomizes the patients to receive either the MTA or the standard therapy within each subgroup. The treatment effect of the MTA can be evaluated by comparing the responses to the different treatments within each marker subgroup[11].

The premise of the MSD is that the biomarker is predictive such that there is a discrepancy of the MTA treatment effect between marker-stratified subgroups. Due to the distinct clinical function of the predictive and prognostic markers, it is reasonable to speculate that the predictive marker used in an MSD has limited prognostic effect. In particular, the response rates for the standard treatment between different marker-stratified subgroups are often close to each other because the targeted biomarker is predictive for the MTA only. For example, the epidermal growth factor receptor (EGFR) is a well known predictive marker for non-small-cell lung cancer (NSCLC) patients[12]. However, recent study has revealed that EGFR contains no prognostic marker effect for NSCLC patients receiving chemotherapy[13]. As a result, the response rates for NSCLC patients receiving non-EGFR-targeted treatment (e.g., chemotherapy) are approximately consistent regardless of patients' EGFR biomarker status. Therefore, a novel design which incorporates these biomarker information into the consideration is expected to improve the efficiency of testing the subgroup treatment effect, which is the primary objective of this paper.

Our study is motivated by a colerectal cancer trial, which is being conducted at the Indiana University Melvin and Bren Simon Cancer Center. The biomarker used in this trial is the KRAS gene mutation. The MTA is a novel KRAS inhibitor and the standard treatment is radiotherapy. The trial use the MSD to evaluate the treatment effect of the

KRAS inhibitor compared with radiotherapy within the marker-positive and marker-negative subgroups, which are stratified by patients' KRAS gene expression level through RNA sequencing. A pilot study indicates that there is no significant difference of the tumor control rates for patients receiving radiotherapy between marker-positive and marker-negative subgroups, which is also confirmed by other studies[14]. Therefore, how to design an efficient marker-stratified trial by integrating this prior information is the challenge of this trial.

In this paper, we propose a Bayesian adaptive MSD to evaluate the subgroup treatment effect of MTAs as well as the predictive marker effect. We develop a Bayesian hierarchical model to capture the similarity of the response rates for patients receiving the standard therapy between different marker-stratified subgroups. We use a tuning parameter within the hierarchical model to represent our prior knowledge regarding the often limited prognostic effect of the biomarker used in the MSD. The magnitude of the tuning parameter is determined by solving a "customized" equation reflecting the reliability of the prior information based on the physician's experience. Therefore, by setting up an informative prior to borrow information across different subgroups, the proposed Bayesian adaptive design achieves higher efficiency while strengthening the individual ethics at the same time. That is, on one hand, the proposed design yields superior power to report promising MTAs and detect significant predictive marker effect with well-controlled type I error rates; one the other hand, the proposed design allocates more patients to more efficacious treatment arms using a response-adaptive randomization algorithm.

The rest of this paper is organized as follows. We propose the Bayesian hierarchical model and introduce how to determine the tuning parameter through a "customized" equation in Section 2. In Section 3, we propose an adaptive MSD based on the hierarchical model. In Section 4, we carry out comprehensive simulation study to investigate the operating characteristics of the proposed design and apply this design to the motivating colerectal trial. We provide a discussion remark in Section 5.

2 Hierarchical Model

Let M be the biomarker indicator with $M = 1$ denoting a marker-positive patient and $M = 0$ denoting a marker-negative patient. Let T be the treatment indicator, with $T = 1$ denoting the MTA and $T = 0$ denoting the standard therapy. We use a binary endpoint Y to indicate whether the patient responds favorably to the received treatment (i.e., $Y = 1$) or not (i.e., $Y = 0$). Let $p_{jk} = \text{pr}(Y = 1|M = j, T = k)$ denote the response probability for patients with marker $M = j$ who receive treatment $T = k$.

Under the MSD, we enroll patients and classify them into different subgroups based on M. Then, within each subgroup, we randomize the patients to receive either $T = 0$ or $T = 1$. The primary objective of the MSD is to test the treatment effects of the MTA within marker-positive and marker-negative subgroups separately[8,9,10,11]. Specifically, let us define $p_{11} - p_{10}$ as the treatment effect of the MTA with respect to the standard therapy in the marker-positive subgroup and $p_{01} - p_{00}$ be the counterpart in the marker-negative subgroup. Under the Bayesian framework, we claim that the MTA is superior to the standard therapy in subgroup j ($j = 0, 1$) if and only if

$$\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D}) > \lambda_1,$$

where $\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D})$ is the posterior probability based on the observed data $\mathcal{D}$, $\delta > 0$ is the minimal meaningful margin for treatment effect and λ_1 is the lower bound cutoff to claim superiority.

Let us focus on the standard therapy arm first. According to our definition, $p_{10} - p_{00}$ is the prognostic effect of biomarker M as it measures the marginal marker effect in the absence of the MTA. Because M is a predictive marker under MSD, its prognostic marker effect can be limited. If this is true, it is reasonable to assume that there is little difference

between p_{10} and p_{00}. In other words, the similarity between these two probabilities should be high. To capture this similarity, we propose a Bayesian beta-binomial hierarchical model. Specifically, let us define n_{jk} as the number of patients with $M = j$ receiving treatment $T = k$ $(j, k = 0, 1)$, and define Y_{jk} as the number of patients among these n_{jk} patients who respond favorably to the treatment $(Y = 1)$. After denoting Binom$(\cdot)$ and Beta$(\cdot)$ be the density functions of the binomial and beta distributions, we model the response rates p_{10} and p_{00} as

$$\begin{aligned} & Y_{10} \sim \text{Binom}(n_{10}, p_{10}); Y_{00} \sim \text{Binom}(n_{00}, p_{00}) \\ & p_{10} \sim \text{Beta}_{p_{10}}(mq, m(1-q)); p_{00} \sim \text{Beta}_{p_{00}}(mq, m(1-q)) \\ & q \sim \text{Beta}_q(\alpha_0, \beta_0) \end{aligned} \tag{1}$$

where m, α_0 and β_0 are the hyperparameters. Under this parameterization, q can be viewed as the "average" response rate for all the patients receiving the standard therapy $T = 0$ and m can be viewed as the prior prognostic effect which characterizes our prior knowledge regarding the prognostic effect for biomarker M. In other words, m measures how close p_{10} and p_{00} are. A larger m indicates less prognostic marker effect and higher correlation between p_{10} and p_{00} *a priori* and therefore more information can be borrowed across different subgroups.

We propose to customize the prior prognostic effect m according to the reliability of our prior information regarding the magnitude of the prognostic marker effect. In particular, we define $f(p_{10}|q, m)$ and $f(p_{00}|q, m)$ as the density functions for the prior distributions of p_{10} and p_{00} conditional on q and m and define $g(q)$ as the density function for the prior distribution of q. Then, after denoting I$(\cdot)$ as the indicator function and τ as a pre-specified cut-off representing the upper-bound of the prognostic marker effect, we characterize the

reliability of the prior information using the probability

$$\begin{aligned}
\mathrm{pr}(|p_{10}-p_{00}|\le\tau|m) &= \mathrm{E}\Big\{\mathrm{E}\big(\mathrm{I}(|p_{10}-p_{00}|<\tau)|q,m\big)\Big\}\\
&= \mathrm{E}\Big(\int_{\mathrm{I}(|p_{10}-p_{00}|<\tau)} f(p_{10}|q,m)f(p_{00}|q,m)dp_{10}dp_{00}\Big)\\
&= \int g(q)\Big(\int_{\mathrm{I}(|p_{10}-p_{00}|<\tau)} f(p_{10}|q,m)f(p_{00}|q,m)dp_{10}dp_{00}\Big)dq.
\end{aligned}$$

Here the second equation holds because p_{10} and p_{00} are conditionally independent given the value of q and we take the expectation of $\int_{\mathrm{I}(|p_{10}-p_{00}|<\tau)} f(p_{10}|q,m)f(p_{00}|q,m)dp_{10}dp_{00}$ with respect to the distribution of q. Hence, $\mathrm{pr}(|p_{10}-p_{00}|\le\tau|m)$ is a function of the prior prognostic effect m. After consulting with the physicians and eliciting a value γ to this probability, the value of m can be determined by solving the following "customized" equation

$$\mathrm{pr}(|p_{10}-p_{00}|\le\tau|m)=\gamma.$$

In addition to m, the prior distribution of q can also incorporate the physician's professional opinion. However, this prior information is often unavailable and a vague prior can be specified for q (e.g., $\alpha_0=\beta_0=0.5$). In this article, we use an uninformative prior for q and an informative m to reflect the setting that we have certain knowledge about the prognostic marker effect but know little about the response rate, which is a common setting under the MSD. Finally, after determining the prior distribution, we can sample the posterior distribution of p_{10} and p_{00} using Gibbs sampler.

We also need the posterior distributions of p_{11} and p_{01} to test the subgroup treatment effects $p_{11}-p_{10}$ and $p_{01}-p_{00}$. Due to the potential predictive effect of marker M, the values of p_{11} and p_{01} may be remarkably different. Hence, we use a beta-binomial model to

characterize these two probabilities as

$$
\begin{aligned}
&Y_{11} \sim \text{Binom}(n_{11}, p_{11}); Y_{01} \sim \text{Binom}(n_{01}, p_{01}) \\
&p_{11} \sim \text{Beta}_{p_{11}}(\alpha_{11}, \beta_{11}); p_{01} \sim \text{Beta}_{p_{01}}(\alpha_{01}, \beta_{01}),
\end{aligned} \tag{2}
$$

and the posterior distribution can be easily determined following a beta distribution.

Finally, with the posterior distributions of all the response rates at hand, we can test the treatment effect for the MTA. As we mentioned earlier, let $\mathcal{D}$ denote all the observed data, we can use the posterior probabilities $\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D})$ to conduct the MSD and test the subgroup treatment effect of the MTA with $M = j$ for $j = 0, 1$, which will be illustrated in detail in the next section. As the proposed design borrows information between p_{10} and p_{00}, which are the response probabilities of the standard therapy in the marker-positive subgroup and marker-negative subgroup, it will yields more precise posterior estimates of both p_{10} and p_{00} compared with the conventional MSD. Furthermore, since the marker-positive treatment effect $p_{11} - p_{10}$ contains p_{10} and the marker negative treatment effect $p_{01} - p_{00}$ contains p_{00}, the proposed design should be more powerful in detecting the treatment effects in both subgroups.

3 Adaptive Design

We propose a Bayesian adaptive MSD based on the aforementioned posterior probabilities $\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D})$ $(j = 0, 1)$. However, in the beginning of an adaptive design, response-adaptive randomization is difficult due to the lack of data. To alleviate this issue, we take a two-stage randomization scheme. Patients come in cohort during the trial. In stage I, each cohort of patients are equally randomized to receive either the standard therapy $T = 0$

or the MTA $T = 1$ after assessing their biomarker status M with a maximum number of cohorts N_1. The objective of this stage is to collect some preliminary data for response rates to facilitate the adaptive randomization in stage II.

The trial cannot be early terminated in stage I, however, we consider early stopping for futility and/or superiority in stage II. Specifically, we define λ_l as the lower-bound cut-off of early stopping for futility and λ_u as the upper-bound cut-off of early stopping for superiority and use these two cut-offs to monitor the early-stopping of the trial during the interim analysis. Assuming that $l-1$ cohorts of patients have been enrolled in the trial, we assign treatments to the lth cohort of patients as follows.

1. Based on the cumulated data $\mathcal{D}_{l-1}$, we update the posterior probability $\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D}_{l-1})$ for $j = 0, 1$ based on the hierarchical model (1) and beta-binomial model (2).

2. If $\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D}_{l-1}) < \lambda_l$, we claim futility of the MTA and stop enrolling patients with $M = j$. Similarly, if $\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D}_{l-1}) > \lambda_u$, we claim superiority of the MTA and stop enrolling patients with $M = j$. If both subgroups are early stopped, then the whole trial is terminated.

3. If the whole trial is not terminated, we assess the biomarker statuses for the lth cohort of patients. Then, if the subgroup $M = j$ is not early stopped, we randomize the patients with $M = j$ to receive the standard therapy $T = 0$ or the MTA $T = 1$ with respective probabilities $\text{pr}(p_{j1} - p_{j0} \leq \delta|\mathcal{D}_{l-1})$ and $\text{pr}(p_{j1} - p_{j0} > \delta|\mathcal{D}_{l-1})$. Otherwise, if the subgroup $M = j$ is early stopped, the patients with $M = j$ should drop off the trial.

4. We collect the response outcomes for the lth cohort of patients and update the cumulative date from $\mathcal{D}_{l-1}$ to $\mathcal{D}_l$.

5. We repeat steps 1-4 until the trial is terminated or the maximum number of cohorts

N_2 for stage II is reached.

If the maximum number of cohorts N_2 is reached, we need to make a final treatment recommendation at the end of the trial. For this purpose, let us define λ_1 as a pre-specified cut-off and assume that subgroup $M = j$ is not early stopped during the trial. Then, we claim superiority of the MTA for subgroup $M = j$ if $\text{pr}(p_{j1} - p_{j0} > \delta | \mathcal{D}_{N_2}) > \lambda_1$ and claim futility of the MTA for subgroup $M = j$ otherwise. As the proposed design uses hierarchical modeling to borrow information, we refer it as HMSD hereafter.

In addition to testing the treatment effect in different subgroup, the HMSD can also be used to detect the predictive marker effect. Following Zang et al.[15], we define $\zeta = (p_{11} - p_{10}) - (p_{01} - p_{00})$ as the predictive marker effect. Then, at the end of the trial, we claim a significant predictive marker effect if $\text{pr}(\zeta > \delta | \mathcal{D}_{N_2}) > \lambda_2$ and claim no predictive marker effect otherwise where λ_2 is the cut-off for predictive marker effect.

4 Simulation and Application

4.1 Simulation studies

We conduct comprehensive simulation studies to investigate the operating characteristics of the proposed HMSD. We compare HMSD with an adaptive-randomized MSD named AMSD and an equal-randomized MSD named EMSD. Both AMSD and EMSD ignore the possible similarity between p_{00} and p_{10} and use the traditional beta-binomial model (2) to model all the p_{jk} $(j, k = 0, 1)$. The AMSD uses the same response-adaptive scheme as HMSD to assign patients in stage II whereas the EMSD insists on the equal randomization through the trial. The difference between the HMSD and AMSD is that HMSD uses the customized hierarchical model to model p_{00} and p_{10} while AMSD uses the beta-binomial model for all the

response rates. The difference between AMSD and EMSD is that AMSD uses the reponse-adaptive randomization in stage II while EMSD uses the equal randomization through the trial. We customize two HMSD according to different prior information, which are referred to as HMSD1 and HMSD2. In HMSD1 we specify $\text{pr}(|p_{10} - p_{00}| \leq 0.05) = 0.95$, resulting in $m = 450$. In HMSD2 we are less confident about the prior information and set $\text{pr}(|p_{10} - p_{00}| \leq 0.05) = 0.80$, which yields a value of $m = 125$.

Table 1 shows the simulation results, including the power/type I error, the number of patients treated with different therapies within different marker subgroups (n_{jk}), the total number of patients treated in the trial (n), probability of early stopping for futility (ESF) and overall response rate for the trial. We set $N_1 = 5$, $N_2 = 25$ and let patients come in a cohort of size 5. The minimum meaningful marginal to claim superiority of the MTA, δ, was set to 0.05. We also specify $\lambda_l = 0.05$, $\lambda_u = 1$ and $\lambda_1 = 0.95$. That is, we allow early stopping for futility rather than superiority, which is reasonable in practice because when a drug is promising, for the purpose of enhancing the individual ethics of the trial, it is often preferred to allocate more patients to that drug. For other hyperparameters in the prior distributions, we select vague priors by specifying $\alpha_0 = \alpha_{01} = \alpha_{11} = 0.3$ and $\beta_0 = \beta_{01} = \beta_{11} = 0.7$. All the results are based on the average of 10,000 simulated trials. For the MCMC procedure, we burn-in the first 10,00 iterations and draw the posterior samples from the successive 5,000 iterations.

In scenario 1, the response rates for both drugs between different marker subgroups are equal (0.3). Therefore, the MTA is not promising for both subgroups and there is no prognostic marker effect at all. All the designs under comparison control the type I errors well around 5%. The other operating characteristics are also similar. The probabilities of early stopping for futility range from 30% to 35% among different designs, resulting in an average sample size around 120 for all the designs. In scenario 2, the MTA and the standard therapy have the same response rates within the marker-negative subgroup (0.3)

and marker-positive subgroups (0.35), with a limited prognostic marker effect of 0.05. Again, all the designs control the type I error well. In particular, compared with scenario 1, using HMSD in scenario 2 only slightly increases the type I error for the marker-positive subgroup, indicating that the proposed design can bear certain amount of prognostic marker effect.

In scenario 3, the MTA is only promising for the marker-positive subgroup. All the designs control the type I error under 5%. For the power comparison in the marker-positive subgroup, The EMSD and AMSD yield similar power. The proposed HMSD is more powerful than both the EMSD and AMSD. Specifically, compared with AMSD, the power of the HMSD1 and HMSD2 is remarkably 17.1% and 11.7% higher. In terms of response rate comparison, the response-adaptive designs (AMSD and HMSD) achieve about 5% higher response rate than the EMSD. The setting of scenario 4 is similar as scenario 3. The HMSD1 reports the highest power and also the highest response rate, and the HMSD2 is the second best design among the three designs (EMSD, AMSD and HMSD2).

In scenarios 5 and 6, the MTA is preferred for both subgroups. As the AMSD is more powerful than the EMSD, we restrict our power comparison between the HMSD and AMSD. The HMSD1 is more powerful than the AMSD under both scenarios. The power improvement is about 4% in marker-negative subgroup and 21% in marker-positive subgroup under scenario 5, and 4% in marker-negative subgroup and 11% in marker-positive subgroup under scenario 6. HMSD2 is less powerful than HMSD1, but still better than the MSD. The AMSD and HMSD yield similar response rates, which are about 5% and 3% higher than the response rate of EMSD under scenario 5 and scenario 6, respectively. In addition, the EMSD allocates about the same number of patients between different drugs regardless of their different response rates. Oppositely, the AMSD and HMSD allocate more patients to the more promising MTA rather than the standard therapy, which is more ethical because more patients can benefit from the trial.

To sum up, the proposed HMSD can well control the type I error if there is no or limited

prognostic marker effect. For power evaluation, the HMSD is the most powerful design across all the scenarios and the power improvement is substantial within the marker-positive subgroup ($10\% \sim 20\%$). In addition, the HMSD1 is more powerful than the HMSD2 because the former design imposes a more informative prior than the latter design. At the meantime, the HMSD is at least as ethical as the AMSD as they yield similar response rates and patients allocation ratios. Therefore, the HMSD outperforms the existing designs and should be recommended in practice.

In Table 2 we report the effective sample size (ESS)[16] of the proposed prior distributions for p_{10} and p_{00} based on the simulation approximation All the ESS are below 1 for all the scenarios considered in Table 1, which indicates a vague prior for p_{10} and p_{00}. Hence, although the proposed hierarchical model makes a strong assumption on the prognostic marker effect, it is still non-informative for the marginal prior distribution of p_{10} and p_{00} so the data should dominate the posterior estimate.

To demonstrate the benefit of using model (1) to detect the subgroup treatment effect, in Table 3 we report the standard deviations of the posterior distributions of p_{10} and the treatment effect $p_{11} - p_{10}$ in the marker-positive subgroup using the EMSD, AMSD, HMSD1 and HMSD2. We consider scenarios 3, 4, 5 and 6 in Table 1. The posterior distribution of p_{10} in EMSD and AMSD were calculated from the traditional beta-binomial model, whereas the posterior distribution of p_{10} in HMSD1 and HMSD2 were calculated from the hierarchical model (1). As we expect, using the hierarchical model (1) yields more precise posterior estimate of p_{10}. Consequently, the HMSD1 and HMSD2 report smaller standard deviations of the posterior distribution of $p_{11} - p_{10}$ compared with the EMSD and AMSD. Hence, based on Table 1 and 3 we conclude that although the prognostic effect is not directly related to the subgroup treatment effect, incorporating the prognostic marker information into the model can improve of the efficiency of the test in detecting the subgroup treatment effect.

All the simulated trials in Table 1 were not allowed to stop for superiority ($\lambda_u = 1$).

However, by adding a stopping rule for superiority, we may expect an increased power and a reduced sample size under the alternative hypothesis. Therefore, it is of interest to investigate the performances of various designs with different values of λ_u other than 1. Table 4 summarizes the simulation results. We consider scenarios 2, 4 and 6 in Table 1 with $\lambda_u = 1, 0.99$ and 0.97 respectively. First of all, we notice that the type I error increased with decreasing λ_u. However, all the designs still control the type I error around 10%, which is in general acceptable for a phase II clinical trial. Secondly, as we expect, adding a stopping rule for superiority can increase power and reduce the sample size at the same time. We take HMSD1 in scenario 4 as an example. When there is no early stopping for superiority ($\lambda_u = 1$), the power of HMSD1 is 82.9% and the sample size is 131.1. The power increases to 85.5% and furthermore 91.1% and the sample size decreases remarkably to 94.3 and 85.0 when $\lambda_u = 0.99$ and 0.97 respectively. We also notice that the response-adaptive designs (AMSD and HMSD) yield smaller response rate by integrating the early stopping rule for superiority. In the aforementioned example, the response rate for the HMSD1 is 34.6% when $\lambda_u = 1$, and is 5.2% and 5.7% smaller when $\lambda_u = 0.99$ and 0.97, respectively. A reasonable explanation is that due to the early-stopping rule, a certain amount of patients who would benefit from the response-adaptive scheme lose their opportunity because the trial is early terminated, which results in a smaller response rate compared with the counterpart without early-stopping rule under the alternative. Therefore, there is a trade-off by integrating the early-stopping rule for superiority into the MSD. Adding this rule will result in an inflated type I error and a decreased response rate. As a reward, the power performance can be improved and the sample size can reduce dramatically. Also, based on the simulation results in Table 4, the proposed HMSD still outperforms the other designs in terms of power and response rate comparison.

As we mentioned earlier, in addition to the subgroup treatment effect, the proposed design can also be used to detect the predictive marker effect. In Figure (1) we depict the

power of the EMSD, AMSD, HMSD1 and HMSD2 in detecting the predictive marker effect, which is denoted by ζ. We let $(p_{00}, p_{01}, p_{10}) = (0.3, 0.4, 0.3)$ and increase ζ from 0.1 to 0.4. As we expect, the power of the HMSD are consistently higher than those of the EMSD and AMSD in detecting the predictive marker effect. For example, when $\zeta = 0.35$, the power for EMSD and AMSD are around 60% and 50% whereas the power for HMSD is over 90%.

4.2 Trial application

We apply the proposed HMSD to the motivating colerectal cancer trial. In particular, for the trial protocol preparation, we investigate the required sample size to achieve 80% power and 10% type I error to detect the treatment effect of the KRAS inhibitor compared with the standard radiotherapy for the marker-positive patients with KRAS gene mutation. The response outcome of this phase II trial is the local tumor shrinkage, which is dichotomized into a binary indicator. Patients come in cohort with a size of 5. We select $\lambda_l = 0.05$, $\lambda = 0.95$ and consider two different values of λ_u as 1 and 0.99. According the simulation results in Table 1, these sets of parameters can control the type I error well below 10%. Based on a pilot study and after consulting with the physicians of the trial, we specify $p_{00} = p_{10} = 0.3$, $p_{01} = 0.4$ and $m = 450$ to reflect the PI's confidence about the similarity between p_{00} and p_{10}. The KRAS mutation rate is about 45% for colerectal cancer patients[17].

Table 5 reports the empirical sample size to achieve 80% power for the EMSD, AMSD and HMSD based on 10,000 simulations with different values of p_{11}. The corresponding response rate for each design is also presented. Among all the designs, the AMSD requires the largest sample size to achieve 80% power and the HMSD is the optimal design in terms of sample size determination. In particular, when $p_{11} = 0.6$ and $\lambda_u = 1$, the required sample size for the AMSD is as large as 276. For comparison, under the same setting, the EMSD and HMSD require 56 and 117 less patients to obtain the same power. We also notice that HMSD yields

almost the same response rate as the AMSD, which is 8.3% higher than the EMSD. When the early-stopping rule for superiority is added ($\lambda_u = 0.99$), all the designs yield significantly lower sample size and the HMSD still obtains the minimum, which is 45 lower than the EMSD and 101 lower than the AMSD. The results for $p_{11} = 0.7$ are similar. In general, when comparing HMSD with EMSD, the HMSD significantly reduces the sample size and improves the response rate. Comparing HMSD with AMSD, the sample size reduction is even bigger and these two designs yield comparable response rates. Therefore, the PI of the colerectal cancer trial has determined to use the HMSD for trial conduction.

4.3 Sensitivity analysis

We conduct a series of sensitivity analyses to investigate the operating characteristics of the proposed HMSD with different sample sizes and response rates. Figure 2 depicts the power changing with different sample size. The parameter setting is the same as scenario 4 in Table 1 except that we let the number of cohorts in stage II (N_2) increase from 1 to 25. As shown in Figure 2, the power of the HMSD is consistently higher than that of the EMSD and AMSD. The magnitude of the power improvement enlarges when N_2 increases. Additionally, in Figure 3 we fix $N_2 = 25$ and let p_{11} increases from 0.26 to 0.50. The conclusion for Figure 3 is the same as Figure 2.

Figure 4 presents the response rates of different designs with different combinations of N_2 and p_{11}. When p_{11} is small ($p_{11} = 0.3$), all the designs report similar response rates. When p_{11} and N_2 increase, the response-adaptive designs (AMSD and HMSD) yield higher response rate than the EMSD and the discrepancy among the response-adaptive designs is generally negligible. Based on Figures 2 to 4, we confirm the conclusion from the simulation studies that the proposed HMSD always yield the largest power to detect the subgroup treatment effect while enhancing the individual ethics at the same time.

In all the aforementioned simulation studies, we assume that the prognostic marker effect is at most limited. Although we believe that this assumption is generally reasonable especially for the MSD with MTAs, it is still of interest to investigate the performance of the proposed design when this assumption is violated. For this purpose, in Figure 5 we depict the type I error rates of all the design with substantial prognostic marker effect. In particular, we fix $p_{00} = p_{01} = 0.3$ and let both p_{10} and p_{11} increase from 0.35 to 0.5 with an increment of 0.05. As the proposed HMSD uses the hierarchical model (1) to borrow information between p_{00} and p_{10}, when there is a difference between these two response rates, the corresponding posterior estimates should shrink toward the "middle" and the magnitude of the bias should be enlarged when the discrepancy becomes bigger.

Figure 5 confirms our speculation. For the marker-positive subgroup $M = 1$, as p_{10} increases, the corresponding posterior estimate based on model (1) is more and more underestimated so the null hypothesis becomes easier to be rejected. As a result, we observe inflated type I errors from the HMSD1 and HMSD2 while the EMSD and AMSD still control the type I error around 5%. We also notice that HMSD2 controls type I error better than the HMSD1. That is because, the HMSD2 uses less informative hierarchical model than the HMSD1 so the bias is also alleviated. Nevertheless, we find that even if the difference between p_{00} and p_{10} is as large as 0.1, HMSD1 still controls a type I error rate around 10%, which is in general acceptable for a phase II clinical trial. The reason is that the purpose of a phase II trial is to quickly identify some positive signal of the drug and send it to a large-scale confirmatory trial rather than rigorously controlling the type I error. The type I error rate will be more strictly controlled in the following confirmatory phase III trial. For the marker-negative subgroup $M = 0$, as p_{00} is overestimated and the null hypothesis becomes harder to be rejected, the proposed HMSD controls the type I error well, just as the EMSD and AMSD. We notice that Figure 5 shows different patterns in controlling the type I errors between the marker positive and negative subgroups. The reason is that p_{10} and p_{00} share

information under the proposed design. In Figure 5, p_{10} is substantially greater than p_{00}. Therefore, due to the borrowed information from p_{00}, p_{10} is underestimated. Similarly, p_{00} is overestimated. Consequently, the difference $p_{11} - p_{10}$ increases which results in a inflated type I error for the marker-positive subgroup. On the other hand, the difference $p_{01} - p_{00}$ decreases so the type I error for the marker-negative subgroup is well controlled.

5 Discussion

We are now entering the era of personalized medicine and a lot of biomarker-guided clinical trial designs have been proposed in the literature. Different biomarker-guided clinical trial designs target for different problems. First of all, the enrichment design[18,19] has been developed to detect the treatment effect of MTAs in a marker-specified subgroup. Gao et al.[20] extended the enrichment design by proposing a multistage adaptive design. The multistage adaptive design incorporates sequential interim analysis into the enrichment design and therefore enhances the flexibility and efficiency of the trial. Moreover, after realizing that the population heterogeneous may substantially affect the performance of the enrichment trial, Gao et al.[21] developed a two-stage design. The purpose of the first-stage is to evaluate the performance of the biomarker and to conduct a sample size re-estimation. Then, a second-stage is followed using the updated sample size to achieve desirable type I and type II errors. Zang and Guo[22] investigated similar topics and developed a optimal two-stage enrichment design to handle the biomarker misclassification. The optimal design can maximizes the probability of correctly classifying each patients biomarker status based on the surrogate marker information. Secondly, the marker-strategy design randomizes patients into non-marker-based or marker-based strategy and test the predictive marker effect by comparing the response rates between two strategies[23]. Zang et al.[15] shown that the

between-strategy comparison is problematic in detecting the predictive marker effect and proposed an optimal marker-strategy design which can maximize the power of detecting the predictive marker effect. Lastly, The marker-stratified design (MSD)[8,9,10,11] is more flexible than the other designs and can be used to test the subgroup treatment effect, predictive marker effect as well as the prognostic marker effect.

In this paper, we propose that the biomarker used in the MSD should be predictive rather than prognostic due to the biological mechanism of the MTA used in the trial. Based on this observation, we extend the MSD and develop a novel hierarchical model to fit the response rate and test the subgroup treatment effect as well as the predictive marker effect. The proposed model uses an informative prior to borrow strength across different subgroups and therefore provides a more accurate estimate than the conventional beta-binomial model. The degree of the prior information is determined by solving a "customized" equation reflecting physicians' professional opinion. Furthermore, we provide a Bayesian adaptive MSD based on the proposed hierarchical model to test the subgroup treatment effects of MTA as well as the predictive marker effect. Simulation results and a real trial application show that the proposed design outperforms the existing conventional adaptive design and equal randomization design. One limitation of the proposed design is that the hyperparameter m need to be elicited prior to the trial and cannot be altered once determined. A possible extension is to specify multiple values of m and use the model selection technique such as the Bayesian model selection to determine the optimal value adaptively during the trial. Further research in this area is warranted.

The proposed design is suitable for the trial scenario where there is a substantial amount of uncertainty about the ability of the biomarker to predict the treatment effect and the investigators are interested in testing the treatment effect for different marker-stratified subgroups. However, if the biomarker used in the trial has distinct predictive effect in the sense that only the marker-positive patients can benefit from the trial. Then, it is unethical

to further randomize the marker-negative patients to receive the treatment. For this circumstance, the marker enrichment design which restricts the enrollment to marker-positive patients only can be used[18,19], which is beyond the scope of this paper. In addition, we note that the proposed design makes sense if the investigator are confident that the prognostic effect of the biomarker used in the trial is limited. If the prognostic marker effect is substantial, the traditional MSD rather than the proposed design should be used because the proposed design can result in an inflated type I error, as shown in Figure 5. In summary, the proposed design can be used in the biomarker-guided clinical trials to detect the subgroup treatment effect and the predictive marker effect when the prognostic marker effect is at most limited. However, as the proposed design pre-specifies the magnitude of the prognostic marker effect, it cannot be used to test the prognostic marker effect.

The idea of using Bayesian model to analyze correlated proportions has been intensively studied in the literature. For example, Altham[24] compared different Bayesian models (logistic, random effect, etc.) for correlated proportions. Kateri et al.[25] provided Bayes and empirical Bayes estimates for the cell probabilities of 2×2 contingency table as well as the Bayes factor for testing the equality of correlated proportions. Agresti and Min[26] studied the operating characteristics of Bayesian confidence intervals for the difference of proportions in a frequentist sense. Oleson[27] proposed Bayesian credible intervals for pre-post binomial proportion correct testing. In the field of adaptive clinical trial design, Thall et al.[28] developed a Bayesian logistic hierarchical model for conducting phase II clinical trial in diseases with multiple subtypes. Yuan and Yin[29] proposed a Bayesian beta-binomial hierarchical model to evaluate the efficacy and toxicity responses for phase I/II trial with combined drugs. However, to the best of our knowledge, little relevant research has been dedicated to the biomarker-guided clinical trial evaluating the MTA. Hence, our work has filled the research gap and we wish the proposed method can attract more attention to this active research area.

Acknowledgments

The authors thank two referees for their valuable comments. The research of Yong Zang is partial supported by the design and biostatistics program pilot grant, Indiana CTSI. The research of Chi Zhang was partially supported by Showalter Trust, Indiana CTSI.

Table 1: The operating characteristics of equal-randomized MSD (EMSD), adaptive-randomized MSD (AMSD), adaptive-randomized MSD with hierarchical modeling (HMSD1 with $m = 450$; HMSD2 with $m = 125$). The number of cohorts is $N_1 = 5$ in stage I and $N_2 = 25$ in stage II. The cohort size is 5. $\lambda_l = 0.05$, $\lambda_u = 1$, $\lambda_1 = 0.95$. ESF denotes the probability of early stopping for futility.

	Power(%)				ESF(%)		
Method	$M = 0$	$M = 1$	$(n_{00}, n_{01}, n_{10}, n_{11})$	n	$M = 0$	$M = 1$	response rate(%)
			Scenario 1. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.3, 0.3, 0.3, 0.3)$				
EMSD	1.9	1.7	(29.2,29.4,29.1,29.4)	117.1	34.7	34.7	30.0
AMSD	5.5	5.6	(29.9,30.1,29.5,30.1)	119.6	30.0	31.6	29.8
HMSD1	3.9	4.6	(31.3,27.6,31.7,27.6)	118.2	32.5	32.0	29.8
HMSD2	2.5	2.9	(32.3,27.6,33.1,27.3)	120.3	31.1	30.2	30.0
			Scenario 2. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.3, 0.3, 0.35, 0.35)$				
EMSD	2.1	1.9	(29.1,28.9,30.0,30.1)	118.1	35.5	31.0	32.4
AMSD	5.6	5.5	(30.3,29.7,30.2,30.0)	120.2	31.2	30.3	32.6
HMSD1	2.2	4.9	(31.5,24.3,31.0,29.0)	115.8	37.6	31.6	32.4
HMSD2	1.8	4.1	(33.3,24.5,32.0,29.8)	119.6	35.8	27.2	32.6
			Scenario 3. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.2, 0.1, 0.25, 0.5)$				
EMSD	0.1	55.8	(19.8,19.9,36.4,36.3)	112.4	72.5	4.2	30.2
AMSD	0.1	53.5	(27.1,15.2,16.1,56.4)	114.8	67.8	3.8	34.9
HMSD1	0.1	70.6	(25.7,13.0,14.1,58.7)	111.5	70.0	4.0	35.8
HMSD2	0.3	65.2	(28.1,13.5,15.0,58.0)	114.6	65.0	3.3	35.3
			Scenario 4. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.2, 0.2, 0.2, 0.5)$				
EMSD	1.7	73.5	(28.6,28.6,36.8,37.2)	131.2	35.9	2.1	28.5
AMSD	4.0	66.5	(30.3,27.4,14.1,59.7)	131.5	34.6	2.2	34.1
HMSD1	5.8	82.9	(30.0,27.2,12.2,61.7)	131.1	35.5	2.1	34.6
HMSD2	3.2	79.8	(31.7,25.9,13.1,60.8)	131.5	36.8	1.2	34.2
			Scenario 5. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.15, 0.35, 0.2, 0.4)$				
EMSD	44.3	41.8	(36.2,36.2,36.3,36.1)	144.8	4.3	5.0	27.5
AMSD	45.0	43.4	(18.3,53.4,18.2,53.3)	143.2	6.3	6.9	32.2
HMSD1	48.9	64.9	(19.0,52.5,15.2,56.5)	143.2	6.8	5.9	32.6
HMSD2	45.2	58.2	(19.1,52.7,16.3,56.0)	144.1	6.3	5.1	32.3
			Scenario 6. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.2, 0.3, 0.2, 0.4)$				
EMSD	12.9	40.9	(33.3,33.6,36.1,36.4)	139.4	15.6	4.4	27.5
AMSD	15.5	43.4	(25.2,41.7,18.6,52.6)	138.1	16.0	6.8	30.6
HMSD1	19.6	54.8	(25.9,41.0,18.3,52.8)	138.0	15.8	7.0	30.6
HMSD2	17.1	50.8	(26.7,41.6,18.4,54.1)	140.8	13.4	4.4	30.4

Table 2: The effective sample size (ESS) of the adaptive-randomized MSD with hierarchical modeling (HMSD1 with $m = 450$; HMSD2 with $m = 125$).

	ESS for HMSD1		ESS for HMSD2	
	p_{10}	p_{00}	p_{10}	p_{00}
Sce. 1	0.983	0.998	0.975	0.987
Sce. 2	0.996	0.991	0.988	0.986
Sce. 3	0.985	0.989	0.976	0.979
Sce. 4	0.980	0.983	0.975	0.973
Sce. 5	0.989	0.991	0.980	0.979
Sce. 6	0.983	0.979	0.974	0.972

Table 3: The standard deviations (s.d.) of the posterior distributions of p_{10} and the subgroup treatment effect $p_{11} - p_{10}$ based on equal-randomized MSD (EMSD), adaptive-randomized MSD (AMSD), adaptive-randomized MSD with hierarchical modeling (HMSD1 with $m = 450$; HMSD2 with $m = 125$).

	s.d. of p_{10}				s.d. of $p_{11} - p_{10}$			
	EMSD	AMSD	HMSD1	HMSD2	EMSD	AMSD	HMSD1	HMSD2
Sce. 3	0.07	0.10	0.05	0.06	0.11	0.12	0.08	0.09
Sce. 4	0.07	0.10	0.05	0.06	0.10	0.12	0.07	0.09
Sce. 5	0.07	0.09	0.05	0.06	0.10	0.10	0.07	0.08
Sce. 6	0.07	0.09	0.06	0.06	0.11	0.11	0.09	0.10

Table 4: The operating characteristics of equal-randomized MSD (EMSD), adaptive-randomized MSD (AMSD), adaptive-randomized MSD with hierarchical modeling (HMSD1 with $m = 450$; HMSD2 with $m = 125$) with early stopping for futility (ESF) and superiority (ESS).

		Power(%)				ESF(%)		ESS(%)		Resp.
λ_u	Method	$M=0$	$M=1$	$(n_{00}, n_{01}, n_{10}, n_{11})$	n	$M=0$	$M=1$	$M=0$	$M=1$	rate(%)
				Scenario 2. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.3, 0.3, 0.35, 0.35)$						
1	EMSD	2.1	1.9	(29.1,28.9,30.0,30.1)	118.1	35.5	31.0	0.0	0.0	32.4
	AMSD	5.6	5.5	(30.3,29.7,30.2,30.0)	120.2	31.2	30.3	0.0	0.0	32.6
	HMSD1	2.2	4.9	(31.5,24.3,31.0,29.0)	115.8	37.6	31.6	0.0	0.0	32.4
	HMSD2	1.8	4.1	(33.3,24.5,32.0,29.8)	119.6	35.8	27.2	0.0	0.0	32.6
0.99	EMSD	6.8	6.8	(28.3,28.5,28.2,28.2)	113.2	33.3	35.2	4.0	3.6	32.4
	AMSD	7.5	7.8	(29.0,29.3,29.3,28.7)	116.3	31.0	31.5	3.7	4.3	32.2
	HMSD1	5.0	8.4	(30.9,22.9,30.6,27.4)	111.8	39.1	29.9	4.4	5.9	32.7
	HMSD2	4.0	8.1	(32.7,24.3,31.1,27.8)	115.9	34.1	27.2	3.2	5.6	32.7
0.97	EMSD	9.4	9.0	(26.5,26.7,27.5,27.6)	108.3	34.6	32.7	8.8	8.7	32.4
	AMSD	9.7	10.5	(29.3,26.0,30.7,26.2)	112.2	32.4	29.7	9.0	9.1	32.5
	HMSD1	8.7	10.5	(30.3,21.7,28.3,24.6)	104.9	38.8	31.1	8.0	13.8	32.6
	HMSD2	5.9	10.9	(32.7,23.3,32.0,24.8)	112.8	34.0	26.7	5.3	11.1	32.5
				Scenario 4. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.2, 0.2, 0.2, 0.5)$						
1	EMSD	1.7	73.5	(28.6,28.6,36.8,37.2)	131.2	35.9	2.1	0.0	0.0	28.5
	AMSD	4.0	66.5	(30.3,27.4,14.1,59.7)	131.5	34.6	2.2	0.0	0.0	34.1
	HMSD1	5.8	82.9	(30.0,27.2,12.2,61.7)	131.1	35.5	2.1	0.0	0.0	34.6
	HMSD2	3.2	79.8	(31.7,25.9,13.1,60.8)	131.5	36.8	1.2	0.0	0.0	34.2
0.99	EMSD	3.1	76.4	(28.0,28.2,24.7,24.5)	105.4	36.8	1.5	2.4	61.3	28.4
	AMSD	5.8	70.1	(30.5,27.7,13.3,41.7)	113.2	33.5	2.2	1.9	41.1	29.8
	HMSD1	6.1	85.5	(30.3,23.1,12.2,28.7)	94.3	38.0	2.2	5.2	71.4	29.4
	HMSD2	5.5	81.8	(32.8,24.4,12.4,32.5)	102.1	32.5	1.9	4.2	64.1	29.6
0.97	EMSD	8.2	82.8	(27.3,27.4,19.0,19.0)	92.7	34.6	2.2	7.7	77.7	28.2
	AMSD	8.1	74.2	(28.9,24.1,13.1,29.3)	95.4	38.4	2.1	7.3	67.0	29.0
	HMSD1	9.8	91.1	(29.9,22.1,11.8,21.2)	85.0	36.1	1.4	9.6	86.2	28.9
	HMSD2	9.4	88.1	(31.4,23.8,12.0,23.2)	90.4	31.7	1.9	8.7	81.9	28.5

Table 4: continued

λ_u	Method	Power(%) $M=0$	Power(%) $M=1$	$(n_{00}, n_{01}, n_{10}, n_{11})$	n	ESF(%) $M=0$	ESF(%) $M=1$	ESS(%) $M=0$	ESS(%) $M=1$	Resp. rate(%)
				Scenario 6. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.2, 0.3, 0.2, 0.4)$						
1	EMSD	12.9	40.9	(33.3,33.6,36.1,36.4)	139.4	15.6	4.4	0.0	0.0	27.5
	AMSD	15.5	43.4	(25.2,41.7,18.6,52.6)	138.1	16.0	6.8	0.0	0.0	30.6
	HMSD1	19.6	54.8	(25.9,41.0,18.3,52.8)	138.0	15.8	7.0	0.0	0.0	30.6
	HMSD2	17.1	50.8	(26.7,41.6,18.4,54.1)	140.8	13.4	4.4	0.0	0.0	30.4
0.99	EMSD	19.9	45.9	(31.3,31.3,30.1,29.9)	122.6	12.9	5.0	12.2	31.5	27.5
	AMSD	21.7	43.6	(25.3,38.7,18.9,43.4)	126.3	14.4	6.2	8.9	21.1	29.7
	HMSD1	27.3	60.1	(25.0,32.5,17.4,34.7)	109.6	14.5	6.5	22.9	44.3	29.3
	HMSD2	21.0	53.8	(26.8,34.4,18.2,37.7)	117.1	15.1	5.3	14.6	36.7	29.2
0.97	EMSD	26.2	55.7	(28.1,28.5,24.4,24.4)	105.4	13.6	5.3	24.2	51.8	27.4
	AMSD	26.9	52.9	(24.6,32.8,17.4,32.9)	107.7	14.4	5.1	23.4	46.5	29.2
	HMSD1	36.3	70.3	(24.2,26.8,16.3,25.3)	92.6	15.8	4.8	33.4	65.1	28.6
	HMSD2	28.7	62.8	(25.8,30.3,17.4,28.7)	102.2	14.9	6.0	24.0	57.0	28.6

Table 5: The sample size and response rates for the motivating colerectal cancer trial to achieve a power of 80% for patients with KRAS mutation. $(p_{00}, p_{01}, p_{10}) = (0.3, 0.5, 0.3)$, $\lambda_l = 0.05$ and $\lambda_1 = 0.95$. The first 5 cohorts of patients are equally randomized to each drug.

		Design					
		EMSD		AMSD		HMSD	
p_{11}	λ_u	Sample size	Resp. rate(%)	Sample size	Resp. rate(%)	Sample size	Resp. rate(%)
0.6	1	220	42.3	276	50.6	159	50.5
	0.99	140	42.2	196	48.5	95	47.9
0.7	1	103	44.8	128	51.8	74	51.2
	0.99	94	43.8	94	48.8	52	48.7

Figure 1: Powers of EMSD, AMSD, HMSD1 and HMSD2 to detect the predictive marker effect ζ. $N_1 = 5$, $N_2 = 25$. $\lambda_l = 0.05$, $\lambda_u = 1$, $\lambda_2 = 0.95$.

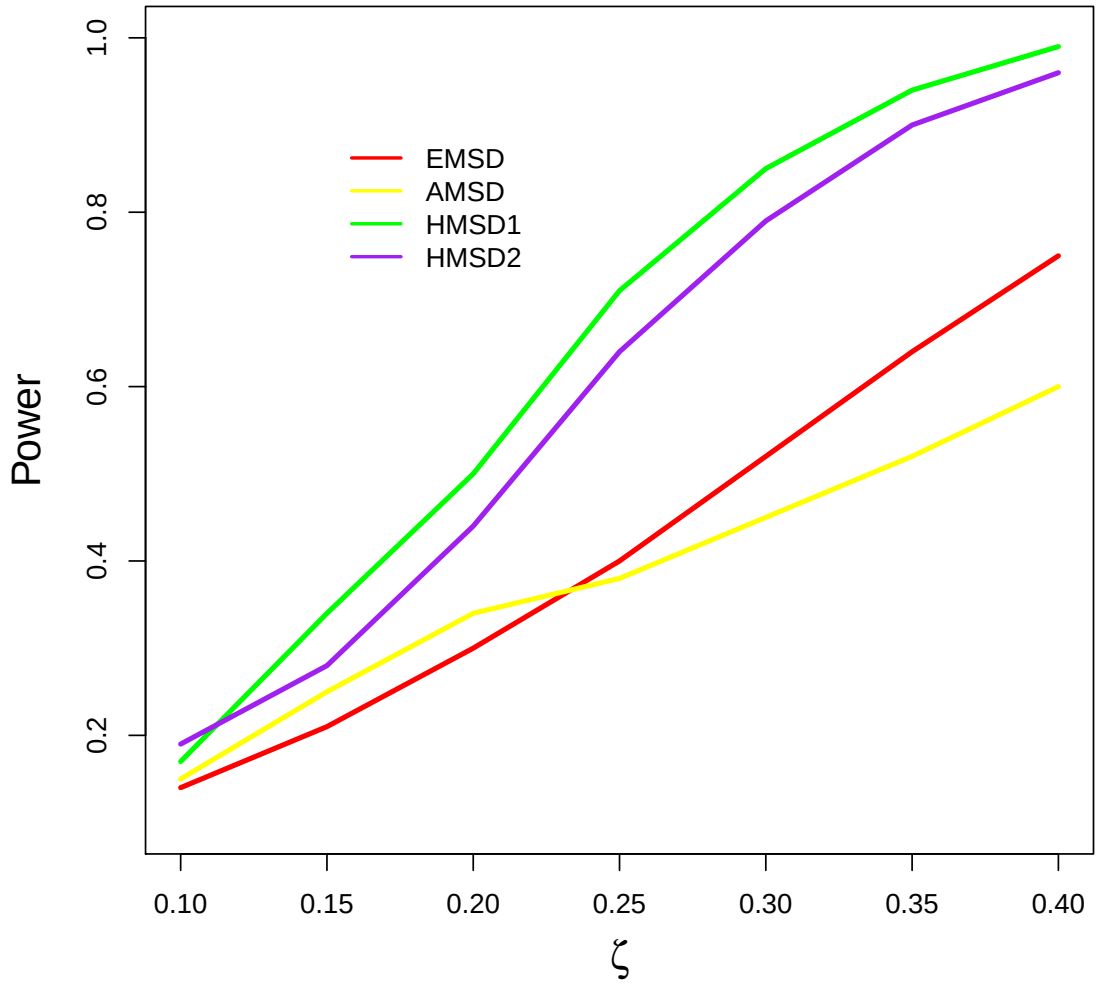


Figure 2: Powers of EMSD, AMSD, HMSD1 and HMSD2 in marker-positive subgroup under different cohort numbers N_2 in stage II. $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.2, 0.2, 0.2, 0.5)$. $\lambda_l = 0.05$, $\lambda_u = 1$, $\lambda_1 = 0.95$.

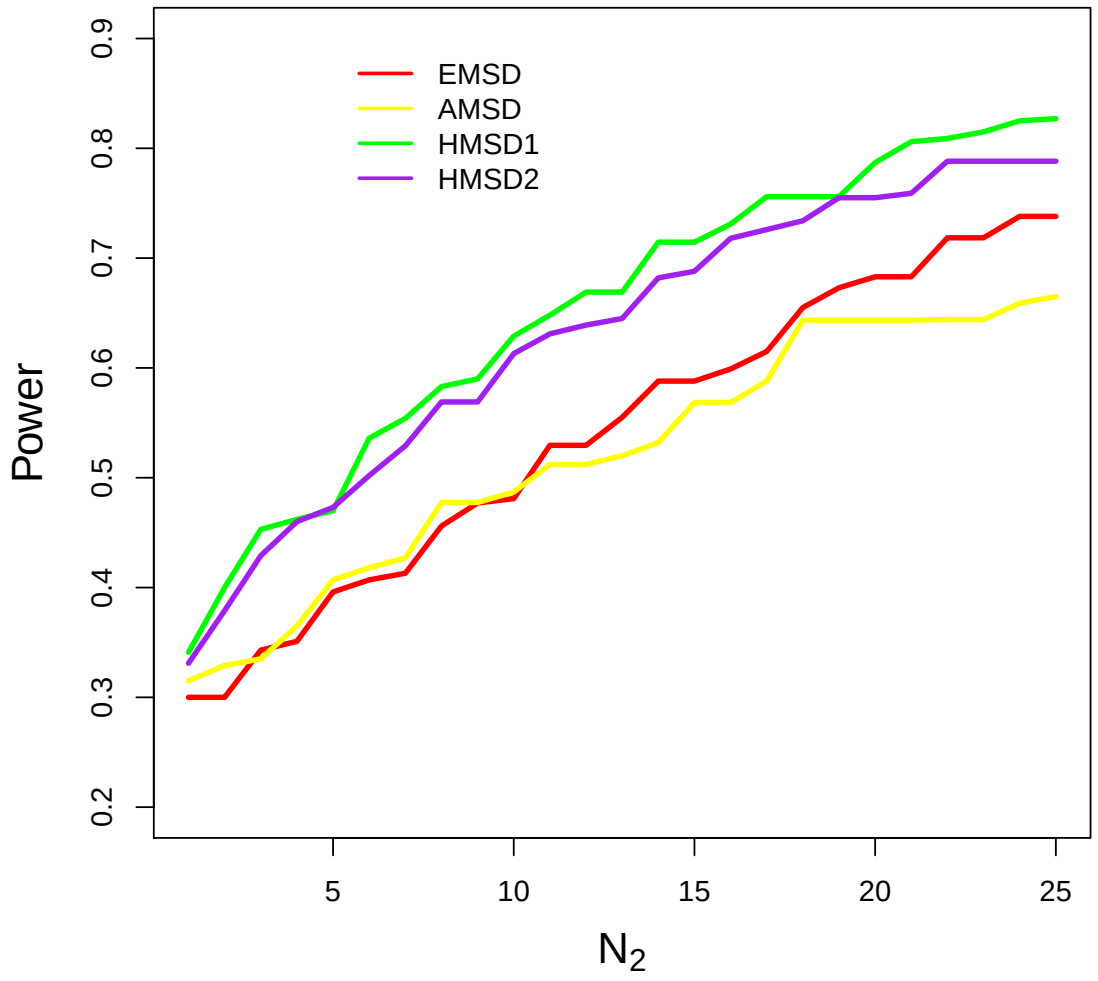


Figure 3: Powers of EMSD, AMSD, HMSD1 and HMSD2 in marker-positive subgroup with different p_{11}. $(p_{00}, p_{01}, p_{10}) = (0.2, 0.2, 0.2)$. $N_1 = 5$, $N_2 = 25$. $\lambda_l = 0.05$, $\lambda_u = 1$, $\lambda_1 = 0.95$.

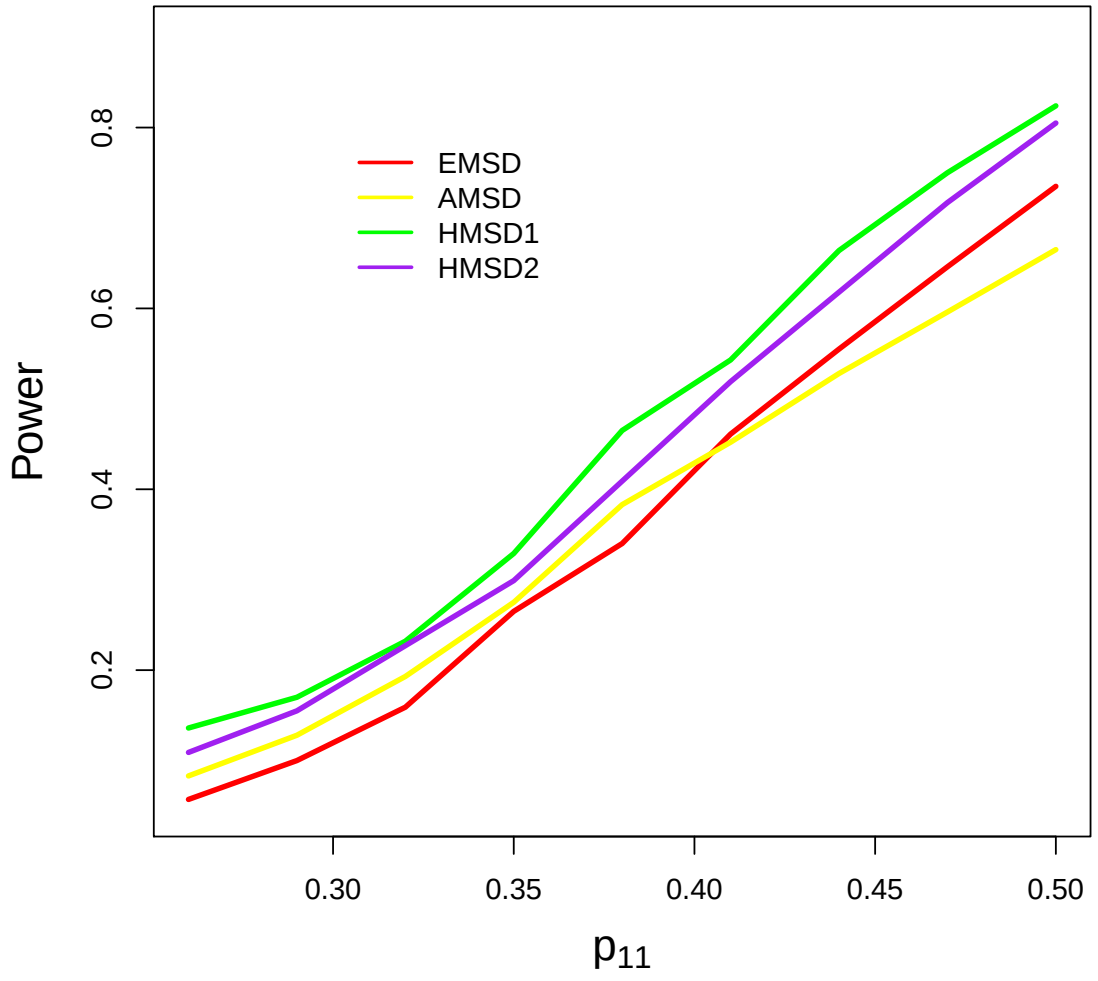


Figure 4: Response rates of EMSD, AMSD, HMSD1 and HMSD2 with different p_{11} and N_2. $(p_{00}, p_{01}, p_{10}) = (0.2, 0.2, 0.2)$. $\lambda_l = 0.05$, $\lambda_u = 1$, $\lambda_1 = 0.95$.

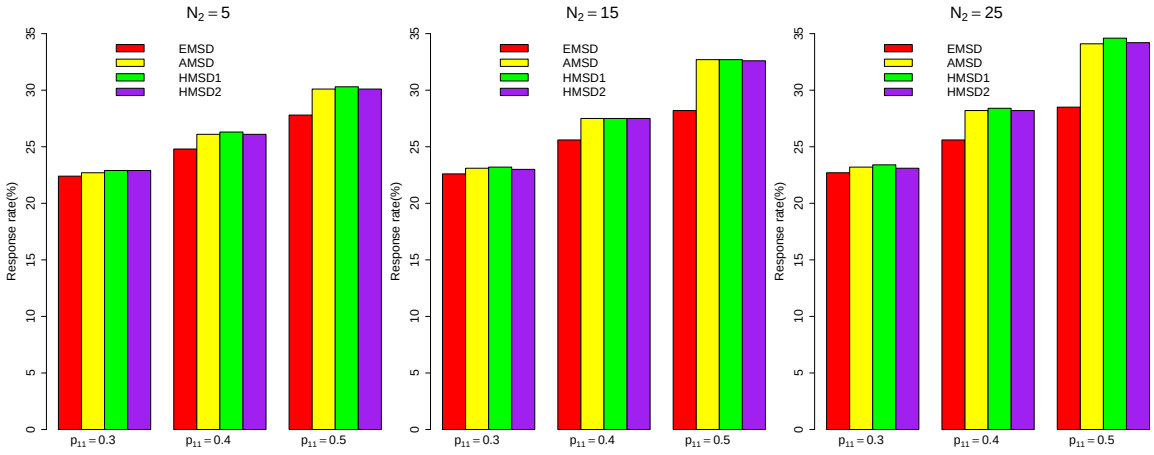
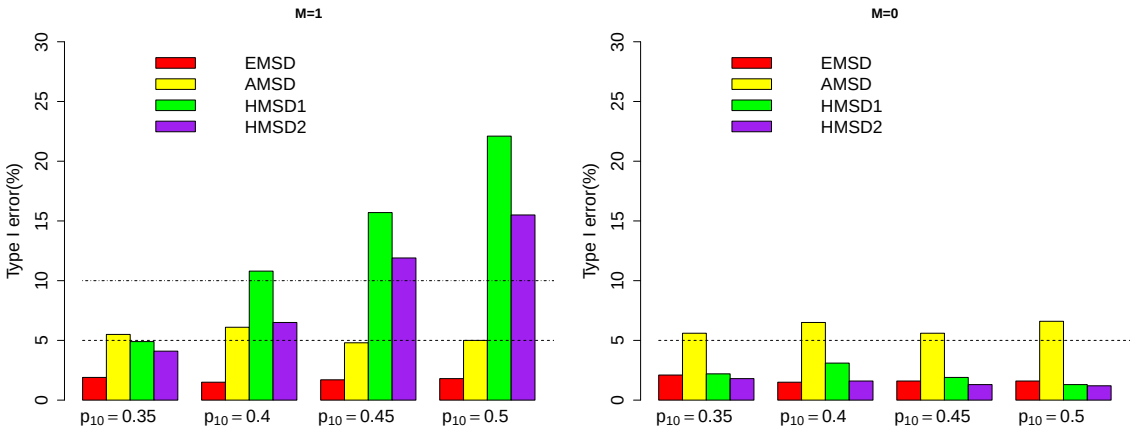


Figure 5: Type I error rates of EMSD, AMSD, HMSD1 and HMSD2 in the presence of prognostic marker effect. $p_{00} = p_{01} = 0.3$, $p_{10} = p_{11}$, $N_1 = 5$, $N_2 = 25$, $\lambda_l = 0.05$, $\lambda_u = 1$, $\lambda_1 = 0.95$.



References

[1] Sawyers, C. Targeted cancer therapy. *Nature* 2004; 432: 294-7.

[2] Sledge GW. What is targeted therapy. *Journal of Clinical Oncology* 2005; 23: 1614-1615.

[3] Wu, HC, Chang, DK and Huang, CT. Targeted therapy for cancer. *Journal of Cancer Molecules* 2006; 2: 5766.

[4] Mandrekar SJ and Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical consideration and practical challenges. *Journal of Clinical Oncology* 2009; 27: 4027-4034.

[5] Sargent DJ, Conley BA, Allegra C and Collette L. Clinical designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 2005; 23: 2020-2027.

[6] Goodsell DS. The molecular Perspective: Tamoxifen and the estrogen receptor. *The Oncologist* 2002; 7: 163-164.

[7] Zang Y, Lee JJ and Yuan Y. Two-stage marker-stratified clinical trial design in the presence of biomarker misclassification. *Journal of Royal Statistical Society: Series C* 2016; 65: 585-601.

[8] Potti A, Mukherjee S, Petersen R, et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *The New England Journal of Medicine* 2006; 355: 570-580.

[9] Kim ES, Hirsh V, Mok T, et al. Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. *Lancet* 2008; 372: 1809-1818.

[10] Wakelee H, Kernstine K, Vokes E, et al. Cooperative group research efforts in lung

cancer 2008: focus on advanced-stage non-small-cell lung cancer. *Clinical Lung Cancer* 2008; 9: 346-351.

[11] Lee JJ, Gu X and Liu S. Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* 2010; 7: 584-596.

[12] Bethune G, Bethune D, Ridgway N and Xu Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *Journal of Thoracic Disease* 2010 2: 48-51.

[13] Kim YT, Seong YW, Jung YJ, Jeon YK, Park IK, Kang CH and Kim JH. The presence of mutations in epidermal growth factor receptor gene is not a prognostic factor for long-term outcome after surgical resection of non-small-cell lung cancer. *Journal of Thoracic Disease* 2013; 8: 171-178.

[14] Dinu D, Dobre M, Panaitescu E, et al. Prognostic significance of KRAS gene mutations in colorectal cancer-preliminary study. *Journal of Medicine and Life* 2014; 7: 581-587.

[15] Zang Y, Liu S and Yuan Y. Optimal marker-strategy clinical trial design to detect predictive markers for targeted therapy. *Biostatistics* 2016; 17: 549-560.

[16] Morita S, Thall PF and Muller P. Determining the effective sample size of a parametric prior. *Biometrics* 2008; 64: 595-602.

[17] Tan C and Du X. KRAS mutation testing in metastatic colorectal cancer. *World Journal of Gastroenterology* 2012; 18: 5171-5180.

[18] Simon R and Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 2004; 10: 6759-6763.

[19] Maitournam A and Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 2005; 24: 329-339.

[20] Gao Z, Roy A and Tan M. Multistage adaptive biomarker-directed targeted design for randomized clinical trials. *Contemporary Clinical Trial* 2015; 42: 119-131.

[21] Gao Z, Roy A and Tan M. A two-stage adaptive targeted clinical trial design for biomarker performance based sample size re-estimation. *Statistics in Bioscience* 2016; 8: 66-76.

[22] Zang Y and Guo B. Optimal two-stage enrichment design correcting for biomarker misclassification. *Statistical Method in Medical Research* 2018; 27: 35-47.

[23] Sargent DJ, Conley BA, Allegra C and Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 2005; 23: 20202027.

[24] Altham PME. The analysis of matched proportions. *Biometrika* 1971; 58: 561-576.

[25] Kateri M, Papaioannou T and Dellaportas P. Bayesian analysis of correlated proportions. *Sankhya: The Indian Journal of Statistics, Series B* 2001; 63: 270-285.

[26] Agresti A and Min Y. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency table. *Biometrics* 2005; 61: 515-523.

[27] Oleson JJ. Bayesian credible intervals for binomial proportions in a single patient trial. *Statistical Method in Medical Research* 2010; 19: 559-574.

[28] Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LH and Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* 2003; 22: 763-780.

[29] Yuan Y and Yin G. Bayesian phase I/II adaptively randomized oncology trials with combined drugs. *Annals of Applied Statistics* 2011; 5: 924-942.